Report Documentation Page

Form Approved
OMB No. 0704-0188

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
28-12-2000	final report	Feb. 1, 1998 to Sept. 30, 2000

4. TITLE AND SUBTITLE

Question-driven Explanatory Reasoning about Devices that Malfunction

5a. CONTRACT NUMBER

5b. GRANT NUMBER
N00014-98-1-0331

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Graesser, Arthur C.

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Department of Psychology
202 Psychology Building
The University of Memphis
Memphis, TN 38152-3230

8. PERFORMING ORGANIZATION REPORT NUMBER
N00014-98-1-0331

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Office of Naval Research
ONR 252
Ballston Centre Tower One
800 North Quincy Street
Arlington, VA 22217-5660

10. SPONSOR/MONITOR'S ACRONYM(S)
ONR

11. SPONSORING/MONITORING AGENCY REPORT NUMBER
N00014-98-1-0331

12. DISTRIBUTION AVAILABILITY STATEMENT
UU

DISTRIBUTION STATEMENT A
Approved for Public Release
Distribution Unlimited

20010103 007

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The process of personnel selection and assignment involves appropriate matches between the abilities of personnel and the jobs assigned to them. For some jobs, personnel need to be selected and trained on the basis of how well they can operate, repair, and maintain particular devices. We have recently discovered two quick and valid methods of determining whether a person has a deep understanding of a mechanical or electronic device. One method involves question asking, the other eye movements. Regarding question asking, we present a breakdown scenario (e.g., *the key turns but the bolt doesn't move*, in the context of a cylinder lock) and observe the quality of the questions that participants ask about causes of the malfunction. Regarding eye tracking, we present the breakdown and observe whether deep comprehenders were more likely to fixate on likely damaged components that explain the breakdown. This research tested a cognitive model of question asking (called PREG).

15. SUBJECT TERMS

personnel selection and classification, question asking, eye tracking, device comprehension, discourse processing, text comprehension

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Arthur C. Graesser
U	U	U	UU		19b. TELEPHONE NUMBER (Include area code) 901-678-2742

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI-Std Z39-18

FINAL REPORT OF OFFICE OF NAVAL RESEARCH
GRANT N00014-98-1-0331
Question-driven Explanatory Reasoning about Devices that Malfunction
For the period of February 1, 1998 to September 30, 2000

Abstract

The process of personnel selection and assignment involves appropriate matches between the abilities of personnel and the jobs assigned to them. For some jobs, personnel need to be selected and trained on the basis of how well they can operate, repair, and maintain particular devices. We have recently discovered two quick and valid methods of determining whether a person has a deep understanding of a mechanical or electronic device. One method involves question asking, the other eye movements. Regarding question asking, we present a breakdown scenario (e.g., *the key turns but the bolt doesn't move*, in the context of a cylinder lock) and observe the quality of the questions that participants ask about causes of the malfunction. Regarding eye tracking, we present the breakdown and observe whether deep comprehenders were more likely to fixate on likely damaged components that explain the breakdown. This research developed and tested a cognitive model of question asking (called PREG).

1. Introduction

Questions are at the heart of virtually any task that an adult performs. It could be argued that any given task can be decomposed into a set of questions that a person asks and answers. For example, when a sailor in the Navy encounters a device that malfunctions, the relevant questions are "What's wrong?" and ""How can it be fixed?". When an officer reads a technical document, the relevant questions are "Why is this important?" and "What should I do about it, if anything?". When a young adult reads Navy recruiting material, the relevant questions are "What's interesting?", "Do I want to join?", and "What are the perks?". The cognitive mechanisms that trigger question asking and exploration patterns need to be understood in order to optimize the design of artifacts, whether they be text, visual displays, mechanical devices, electronic equipment, or telecommunication systems.

One of the goals of this ONR project was to develop and test a cognitive computational model of question asking. We developed such a model, called PREG, which means question in the Spanish language (Graesser, Olde, Pomeroy, Whitten, Lu, & Craig, in press; Otero & Graesser,in press). According to the PREG model, cognitive disequilibrium drives the asking of genuine information-seeking questions (Berlyne, 1960; Chinn & Brewer, 1993; Collins, 1988; Festinger, 1957; Flammer, 1981; Graesser, Baggett, & Williams, 1996; Graesser & McMahen, 1993; Graesser & Person, 1994; Schank, 1999). Questions are asked when individuals are confronted with obstacles to goals, anomalous events, contradictions, discrepancies, salient contrasts, obvious gaps in knowledge, expectation violations, and decisions that require discrimination among equally attractive alternatives. The answers to such questions are expected to restore equilibrium and homeostasis. Otero and Graesser (in press) developed a set of production rules that specify the categories of questions that are asked under particular conditions (i.e., content features of text and knowledge states of individuals). It often takes a large among of knowledge to identify such clashes and gaps in knowledge. Miyake and Norman (1979) presented the argument over 20 years ago that "to ask a question, one must know enough to know what is not known."

Questions that tap explanatory reasoning are particularly diagnostic of deep comprehension. When considering equipment, explanations are needed when devices break down, faults are diagnosed, and devices are repaired. The person responsible for a broken piece of equipment needs to construct explanations in the form of causal networks, goal-plan-action hierarchies, and logical justifications. It is

well documented that the construction of explanations is a robust predictor of an adult's ability to learn technical material from written texts (Chi, deLeeuw, Chiu, & LaVancher, 1994; Cote, Goldman, & Saul, 1998; Graesser, VanLehn, Rose, Jordan, & Harter, in press; VanLehn, Jones, & Chi, 1992).

Question asking tasks have the potential for improving the accuracy of personnel selection and classification. For example, a sailor would ideally be assigned to be a locksmith if the sailor has deep knowledge that explains lock mechanisms, but not if the sailor merely knows the jargon. But how does one know whether a sailor has the talent and the deep knowledge for a task? We know that we will not get much useful information by simply asking the sailor (e.g., "How good are you in operating a lock?"). There are serious limitations in the metacognitive abilities of adults in monitoring the accuracy of their own comprehension (Hacker, Dunlosky, & Graesser, 1998). We know that we will not get much useful information by testing the sailor on inert shallow knowledge, such as a test of vocabulary and technical jargon (e.g., "What is cam?"). Shallow knowledge is hardly a substitute for deep knowledge. We know that it would be impractical to spend several years developing a fully validated, reliable, psychometric test on each device in the military. The device would be outdated by the time the psychometric test was finished.

The present project investigated the questions that college students ask when an everyday device malfunctions. After reading about a device (e.g., cylinder lock, dishwasher), the participants subsequently received scenarios in which the device breaks down (e.g., *the key turns but the bolt doesn't move*, in the context of a cylinder lock) and they generated questions about the malfunction. Eye tracking data were also collected during question asking in one of the empirical studies. There are two straightforward predictions of the PREG model. First, those participants who have a deep understanding of the device should ask good questions that converge on faults. Second, the eye movements of deep comprehenders should converge on likely faults that explain the breakdown.

The remainder of this final report is divided into five parts. Section 2 identifies the levels of knowledge representation that are potentially constructed during the comprehension of an illustrated text. Section 3 reports an empirical study that tests the prediction that deep comprehenders ask good questions when devices break down. Second 4 reports a study that collects eye tracking data and tests the prediction that good comprehenders tend to focus on faults in breakdown scenarios. Section 5 describes the PREG model and some of its more subtle predictions. Section 6 briefly identifies some of the practical implications of this research.

2. Comprehending Illustrated Texts at Different Levels of Representation

Adults occasionally read illustrated texts that describe the mechanisms of an everyday device, such as the cylinder lock depicted in Figure 1. As the printed text is read, there is an attempt to decipher and integrate the components, labels, spatial relations, and arrows in the pictures (Hegarty & Just, 1993). An ideal comprehender attempts to understand the mechanism at a deep level.

Discourse psychologists and cognitive scientists have identified the different levels of representation that are affiliated with shallow versus deep comprehension (Britton & Graesser, 1996; Gentner & Stevens, 1983; Graesser, Millis, & Zwaan, 1997; Kieras & Bovair, 1984; Kintsch, 1998). The most shallow level is the surface code, which preserves the exact wording and syntax of the explicit verbal material. When considering the visual modality, it preserves the low-level lines, angles, sizes, shapes, and textures of the pictures. At an intermediate level, there is a propositional representation that captures the meaning of the explicit text and the pictures. At the deepest level, there is the mental model of what the text is about. For everyday devices, this would include: the components of the electronic or mechanical system, the spatial arrangement of components, the causal chain of events when the system successfully unfolds, the

mechanisms that explain each causal step, the functions of the device and device components, and the plans of agents who manipulate the system for various purposes. Quite clearly, a rich set of knowledge structures get constructed when an adult comprehends a device at a deep level.

Researchers in discourse psychology and artificial intelligence have developed theories that specify how to organize and represent world knowledge. This knowledge consists of component hierarchies, spatial layouts, causal mechanisms, goal-driven procedures, and various other types of knowledge (Graesser & Clark, 1985; Graesser, Gordon, & Brainerd, 1992; Kintsch, 1998; Lehmann, 1992; Lenat, 1995; Perfetti, Britt, & Georgi, 1995; Trabasso & van den Broek, 1985; Schank, Kass, & Riesbeck, 1994). We believe that a detailed analysis of these mental models is needed to gain a sophisticated theoretical understanding of deep comprehension.

For example, Figure 2 presents a portion of the knowledge structure that depicts the explicit information in an illustrated text about a cylinder lock. This is based on the conceptual graph structure representations developed by Graesser (Baggett & Graesser, 1995; Graesser & Clark, 1985; Graesser et al., 1992; Graesser, Wiemer-Hastings, & Wiemer-Hastings, in press). The composition of these conceptual graph structures is not arbitrary, but is based on formal and conceptual constraints that have been studied for three decades. For example, the categories of nodes and arcs are functionally adequate for implementing models of question answering and questions asking that have been supported in experiments on adults (Graesser & Hemphill, 1991; Graesser, Lang, & Roberts, 1991; Otero & Graesser, in press). At the center, there is a causal chain of events that unfold when the key successfully unlocks the door. At the left, there are nodes that capture part of the spatial composition of the system that causally enables the events. At the right, there are goals of the agent (person) who interacts with the lock. Some nodes need to be inferred to make sure the graph is coherently organized. We believe that knowledge structures such as these are constructed in the mind during the process of comprehending an illustrated text and later reflecting on the content while solving problems.

A few words should be devoted to the terminology that is associated with conceptual graph structures. A conceptual graph structure consists of a set of nodes that are connected by a set of directed, categorized arcs. The nodes are concepts and proposition-like descriptions that refer to either text constituents or visual-spatial aspects of the pictures. Thus, there is a picture description language that can translate most aspects of a picture into a structured description. For example, the concepts associated with a cylinder lock would include the following noun referents: *lock*, *pins*, *cam*, *spring*, *rod*, *bolt*, and so on. The proposition-like descriptions are categorized as States (*The rod is next to the cam*), Events (*The cam rotates*), Goals (*Turn the key*), or some other ontological category. The nodes are connected by different categories of arcs that specify Causality (C) and enablement, Reasons (R) for generating goals, and Outcomes (O) of goals. The complete representational system has 22 basic arc categories (Graesser, Wiemer-Hastings, & Wiemer-Hastings, in press), but it is beyond the scope of this report to get into all of the details about the representational system. Most arc categories are directed, with a source node and an end node. For example, in the case of Cause arcs, the cause must temporally precede the effect so the arrow points from the source node (*The cylinder rotates*) to the end node (*The cam rotates*): (Node-5) –C→ (Node 6) in Figure 2.

Some pieces of knowledge are depicted in the picture, some in the text, and some in both. For example, consider the nodes in Figure 2. The nodes that are depicted strictly in the picture are Event 6 and States 1, 2, 3, and 4. The nodes that are expressed only in the text are Events 1, 2, 3, and 5 and Goals 1 and 2. The nodes that are captured in both the picture and text are Events 4, 7, and 8. One might expect a reader with high verbal aptitude to focus on the nodes captured in the text, whereas readers with high visual-spatial

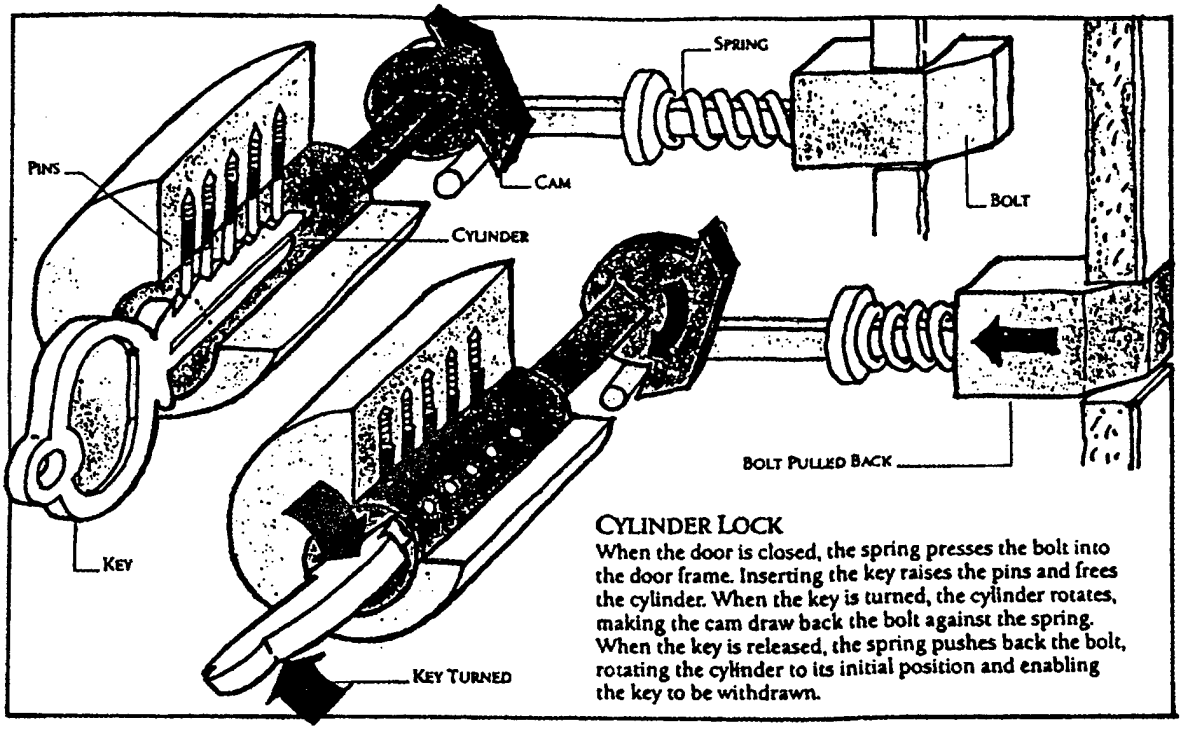

Figure 1: Example illustrated text describing a cylinder lock. (From *The Way Things Work* by David Macaulay. Compilation copyright (c) Dorling Kindersley Ltd., London. Illustration copyright (c) 1988 David Macaulay. Text copyright (c) 1988 David Macaulay, Neil Ardley. Reprinted by permission of Houghton Mifflin Company. All rights reserved.)

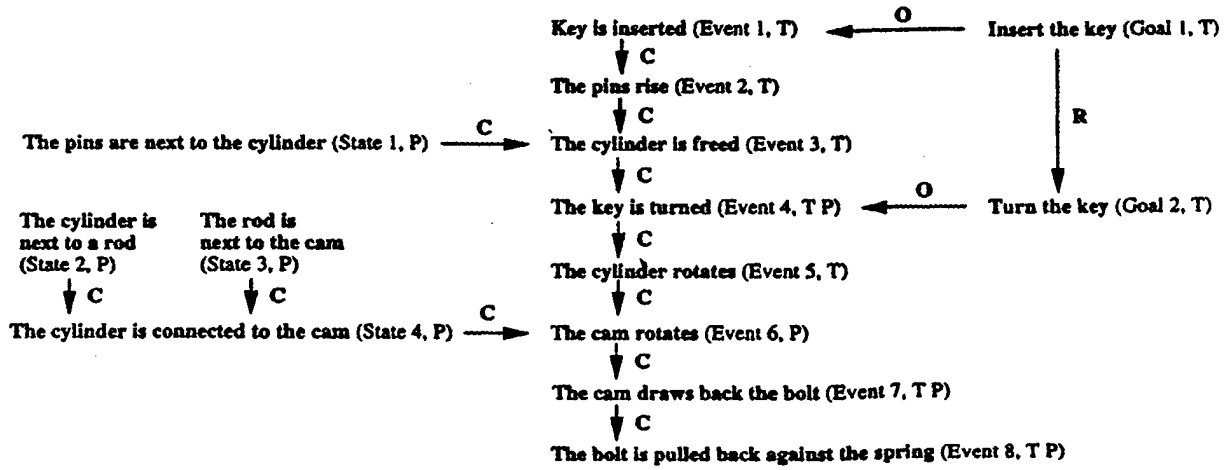

Figure 2: A portion of the cylinder lock materials represented as a conceptual graph structure. The arc categories incluc consequence (C), reason (R), and outcome (O). Each node is labeled as being depicted in the picture alone (P), the text alor (T), or both text and picture (T P).

ability to focus on the nodes captured in the picture. There is some evidence that the text dominates the reading process when college students read illustrated texts for comprehension (Baggett & Graesser, 1995; Hegarty & Just, 1993). In Hegarty and Just's research on eye tracking, for example, the text drives eye movements. The eye movements drift to the pictures primarily when there is a term or proposition in the text that is unknown, unresolved, or confusing. In Baggett and Graesser's (1995) research on question answering, the text-based nodes occur in answers to deep-reasoning questions *(why, how, what-if, what-if-not)* with a higher incidence than picture-based nodes. They also reported that nodes appear in the answer with a very high likelihood if they are depicted in both text and picture. This latter finding would be predicted by a dual-code theory of multimedia processing (Mayer, 1997; Mayer & Sims, 1994).

The practical importance of achieving deep comprehension is perfectly obvious. In the arena of education and training, one of the important missions is to enhance deep understanding of the domain knowledge in science, mathematics, history, and other areas. It is not enough to impart shallow, inert knowledge; the students also need to acquire deep knowledge that can be actively put into practice in practical applications and that can help solve difficult problems. In the arena of selection and classification, there is the need to assign sailors to tasks and duties that are suited to them. Ideally, a sailor is assigned to be a locksmith, for example, if the sailor has deep knowledge about the causal mechanisms that explain locks, but not if the sailor merely knows the jargon and marketing hype.

In this project we investigated some methods that provide a quick litmus test of the extent to which a college student has achieved a deep comprehension of a device. One litmus test is based on two central assumptions about deep comprehension. First, understanding is manifested when the device breaks down, not when it is running smoothly. Explanatory reasoning is not particularly critical when a device is running smoothly and the human operator has minimal interactions with the device. Explanations are needed when devices break down, faults are diagnosed, and devices are repaired. Second, understanding is manifested in the quality of questions that an adult asks while they reflect on a breakdown scenario. That is, deep comprehenders ask good questions that converge on the faults that explain the breakdown. From the standpoint of the conceptual graph structures, a deep comprehender asks questions that converge on those nodes in the structure that are plausible causes of the breakdown. In summary, a good litmus test of deep comprehension for a device lies in the quality of questions that are asked in the context of a breakdown scenario.

Consider an example of a breakdown scenario that might occur in the context of a cylinder lock. Suppose that an adult named Jack is confronted with the following breakdown:

BREAKDOWN SCENARIO: The key turns, but the bolt doesn't move.

That is, Jack moves the key and it has no trouble turning, but unfortunately the bolt does not move back and forth (see Figure 1). Jack may explicitly and overtly ask questions about potentially causes of the malfunctions. Alternatively, the questions might merely pop into his mind, with varying degrees of precision, vividness, and completeness. The questions might be implicit, but be manifested by the actions that Jack performs or the assertions that Jack expresses overtly. If Jack wonders whether the cam is broken, one of the following events might occur:

(1) Jack asks: "Is the cam rotating?"
(2) Jack says: "The cam might have trouble rotating."
(3) Jack looks at the cam.
(4) Jack observes the cam as he moves the key.

All of these acts are counted as question asking in the sense that they reflect the process of inquiry, uncertainty, and curiosity. Jack wonders whether the cam is causing the breakdown so he performs some

physical act, perceptual act, or verbal act to reduce the uncertainty and satisfy his curiosity. It is unimportant whether the question is manifested perceptually, physically, verbally, or in an interrogative syntactic form. Question asking emerges in any of these forms.

The question that Jack asks is a good one because it identifies a likely fault that would explain the unmoving bolt. A good question converges on likely faults of the breakdown. In the example breakdown, the likely faults would correspond to two nodes in Figure 2: Event 6 (*The cam rotates*) and Event 7 (*The cam draws back the bolt*). The remaining 12 nodes in Figure 2 would not be the locus of likely faults. There are a large number of questions that would be bad questions because they would not explain the malfunction. For example, the following questions are low quality because they fail to account for the breakdown.

> (5) "Are the pins rising?" The pins would have to be rising because the key is inserted and successfully turning.
> (6) "Is the right key?" It is very likely the right key because the key is successfully inserted and turning.
> (7) "Is the spring broken?" The spring can assist the bolt in moving, but it can not prevent the bolt from moving.

Deep comprehension would be manifested by the questions reflected in 1-4, whereas shallow questions would be reflected in questions 5-7.

One obvious question to ask is why question asking provides such a good litmus test of deep comprehension. Why wouldn't a "think aloud" task provide a more reliable window into deep comprehension, as would be advocated by Ericsson and Simon (1993) and others? Our argument is that the mechanisms of question asking are particularly tailored to breakdown scenarios. Cognitive disequilibrium drives the asking of genuine, information seeking questions, as predicted by the PREG model (Graesser, Olde, Pomeroy, et al., in press; Otero & Graesser, in press) and the available literature on question asking (see Introduction).

3. Asking Questions when Devices Break Down: An Empirical Study

We conducted a study that tested the prediction that a good litmus test of deep comprehension consists of the questions that are asked in the context of a breakdown scenario. College students at the University of Memphis (N = 108) first read an illustrated text, then were given a breakdown scenario, and then generated questions in writing. The questions that participants ask should have higher quality if they have deeper comprehension of the device. A question was scored as a high in quality if it referred to a likely fault that explained the breakdown. More specifically, some of the nodes in the knowledge structure for a device were likely faults; high quality questions matched or directly referred to those nodes. After completing the question asking task, the college students were given an objective comprehension test on the devices. The obvious prediction is that performance on this device comprehension test should positively correlate with the quality of the questions that get asked. The participants also completed a battery of tests of cognitive ability and personality. We investigated how well these other measures of individual differences compared to the quality of questions in predicting device comprehension.

Methods

Illustrated texts and tasks. The participants read 6 illustrated texts on everyday devices: a cylinder lock, an electronic bell, a car temperature gauge, a clutch, a toaster, and a dishwasher. The device mechanisms were extracted from Macaulay's book with illustrated texts, ***The Way Things Work*** (Macaulay, 1988). After reading about each device, the participants subsequently received scenarios in

which the device breaks down (e.g., *The key turns, but the bolt does not move* in the context of the cylinder lock). During this time, the participants were asked either to "think aloud in writing" (which we will call the Write Aloud task) or to generate questions in writing (Question Asking task) for three minutes. The participants typically reflected on how to diagnose and repair the malfunctions during the Write Aloud and Question Asking tasks. The Write Aloud task was completed for three devices prior to the Question Asking task, which in turn was completed for three other devices. The assignment of devices to conditions and test order was counterbalanced across 108 college students at the University of Memphis.

Device comprehension test. After providing the Question Asking and Write Aloud protocols for all 6 devices, the participants completed an objective test on their understanding of the devices. This consisted of six 3-alternative, forced-choice questions about each device (36 total questions across the 6 devices). There were 4 test questions per device that tapped explicit information and 2 questions that tapped inferences. Examples of such questions are provided below.

EXPLICIT: What action by a person causes the pins to rise?
(a) the key is inserted (correct answer)
(b) the key is removed
(c) the key is turned

INFERENCE: What happens to the pins when the key is turned to unlock the door?
(a) they rise
(b) they drop
(c) they remain stationary (correct answer)

The device comprehension scores could vary from 0 to 36. A score of 12 would be chance performance if there were no sophisticated guessing or auxiliary background knowledge.

The device comprehension test was defined as the gold standard for deep comprehension. The questions were generated systematically by adopting a theoretical foundation in qualitative physics (Forbus, 1984). Suppose there is a set of N component nodes in a system, which are connected by a network of -, +, and 0 causal relations. If node C is affected in some fashion (e.g., increased input, broken, initiated, rotated), how would it propagate its effects on the other nodes in the system (e.g., components X, Y and Z)? There are always 3 alternative answers that reflect the impact on an effected node, such as (a) X increases, (b) X decreases, and (c) X stays the same. Thus, a deep comprehender is able to trace the causal antecedents and causal consequences of an event (Graesser & Bertus, 1998).

Battery of tests of individual differences. Following the objective test of device comprehension, participants completed a battery of tests that measured their cognitive abilities and personality. The tests of cognitive ability included the ASVAB (the Armed Services Vocational Aptitude Battery, Department of Defense, 1983). This test is administered to over 1 million high schools students each year. There were the following subscales on this test: Mechanical comprehension, electronics, general science, auto & shop, mathematics knowledge, arithmetic reasoning, numerical operations, word knowledge, paragraph comprehension, and coding speed. Five composite variables can be derived from the 10 measured variables on the ASVAB: technical knowledge, verbal ability, quantitative ability, speed, and general intelligence (g). Additional tests of cognitive ability included working memory span (LaPointe, & Engle, 1990), spatial reasoning (Bennet, Seashore, & Wesman, 1972), and exposure to print (the author recognition test, Stanovich & Cunningham, 1992).

A number of noncognitive variables were measured. These included age, gender, and scales on a personality test. The personality test is the NEO inventory (Costa & McCrae, 1991), which measures individuals on the "big five" personality factors: neuroticism, extroversion, openness, agreeableness, and conscientiousness. The subscale of openness attempts to capture creativity, which we anticipated might be correlated with question asking. It took approximately 4 hours to complete the battery of tests, which were completed in two sessions on two different days.

Measures of question asking and write aloud. Four measures were scored on the verbal protocols that were collected in the Question Asking and Write Aloud tasks. These are listed and defined below.

Volume of questions. The number of questions that were asked in the Question Answering task.
Question quality. The proportion of questions that referred to a plausible malfunction that explained the breakdown.
Volume of ideas. The number of ideas expressed in the Write Aloud task
Idea quality. The proportion of ideas that referred to a plausible breakdown.

Trained judges segmented the protocols into separate questions or idea units. There was a high reliability in such judgments (.90 or higher between any given pair of judges). Trained judges also determined whether a question or idea matched a fault node. This judgment required more training, but pairs of judges did eventually reach an acceptable level of agreement (.80 or higher in common decisions).

Results and Discussion

Descriptive Statistics. Table 1 presents means and standard deviations for the measures that were collected in this study. This includes the ASVAB scores, spatial reasoning, working memory span, exposure to print, the personality measures (NEO), gender, age, measures of the verbal protocols, and device comprehension scores.

The measures of individual differences presented no surprises when compared to normal college student populations. As would be expected, the general intelligence scores and other subscales were above average compared with the population of high school students who take the ASVAB. The other cognitive measures are not significantly different from the scores for college students that are reported in Bennet et al. (1972) for spatial reasoning, LaPointe and Engle (1990) for working memory span, and Stanovich and Cunningham (1992) for exposure to print. It should be noted that the working memory span measure gives credit for partial answers, as opposed to being a measure that estimates number of chunks. The five personality subscales on the NEO inventory Costa & McCrae, 1991) were all hovering around the population mean of 50. There were more females (62%) than males in the sample, which is consistent with the estimates of college populations in the year 2000 (59% being female).

A number of observations can be made about the measures of the verbal protocols in the Question Asking and Write Aloud tasks. The mean was somewhat higher for the Write Aloud Task than the Question Asking task, but the standard deviations were very close. The quality of questions and ideas was, once again, measured by computing the percentage of verbal units (questions or ideas) that matched one of the nodes in the conceptual graph structure that would explain the breakdown; these are called fault nodes. According to the data, 22% of the questions and 17% of the ideas were high in quality.

Table 1. Descriptive statistics on measures collected in the question asking experiment

MEASURES	DESCRIPTIVE STATISTICS	
	Mean	Standard Deviation
Cognitive Measures		
ASVAB (g)	125.3	19.4
Mechanical Reasoning (MR)	14.7	5.0
Electronics (EL)	11.4	4.0
General Science (GS)	18.6	4.3
Auto & Shop (AS)	12.3	5.5
Mathematics Knowledge (MK)	18.3	4.7
Arithmetic Reasoning (AR)	22.4	5.5
Numerical Operations (NO)	39.4	7.9
Word Knowledge (WK)	29.9	5.1
Paragraph Comprehension (PK)	12.5	3.3
Coding Speed (CS)	58.8	12.3
Spatial Reasoning (SP)	27.3	14.4
Working Memory Span (WM)	33.4	8.6
Exposure to Print (EP)	9.1	6.9
Personality Measures (NEO)		
Neuroticism (N)	49.9	12.3
Extroversion (E)	52.7	12.3
Openness (O)	51.8	11.6
Agreeableness (A)	46.3	13.5
Conscientiousness (C)	47.2	12.3
Demographics Measures		
Gender (GEN, female = 1, male =2)	1.38	.49
Age	24.7	7.7
Verbal Protocol Measures		
Volume of questions (VQ)	3.8	2.0
Quality of questions (QQ)	22.4%	14.7
Volume of Ideas (VI)	5.4	1.8
Quality of ideas (QI)	17.3%	13.4
Device Comprehension Score	23.5	5.3
Number of Participants	108	

The gold standard for measuring deep comprehension was the device comprehension score. The mean score was 23.5 out of 36 questions, so 65% of the 3-alternative, forced-choice questions were answered correctly. The questions that tapped explicit information in the illustrated texts were answered correctly more often than those that required inferences, 71% versus 54%, respectively. We do not distinguish these subclasses of questions in the remainder of the report, however.

Correlations. Table 2 presents correlation coefficients that are relevant to assessments of variables that predict device comprehension scores and to the measures of verbal protocols. This table includes four additional composite measures that are provided by ASVAB:

Technical Knowledge: Mechanical Reasoning, Electronics, General Science, Auto & Shop
Verbal: Word Knowledge, Paragraph Comprehension
Quantitative: Mathematics Knowledge, Arithmetic Reasoning
Speed: Numerical Operations, Coding Speed

Consider first the prediction of device comprehension scores, which are shown in the left column of numbers. When $r = .50$ is adopted as a minimum threshold for a robust correlation, device comprehension scores were predicted by question quality, spatial reasoning, the technical knowledge composite (and each of its component measures), the quantitative composite (and each of its component measures), and ASVAB general intelligence. Several measures are nonsignificant ($r < .20$) so they fail to predict device comprehension scores: volume of questions, volume of ideas, age, most personality measures (N, E, A, C), working memory, and the speed composite measure (and each of its component measures). In between these two thresholds are a number of modest, but significant correlations: Quality of ideas, gender, openness, exposure to print, and the verbal composite measure (and each of its component measures).

According to these results, question quality is a robust predictor of device comprehension scores, on par with psychometric measures that are expected to predict device comprehension. Indeed, the .51 correlation between question quality and device comprehension scores is not significantly different than the robust noncomposite measures of ASVAB (which vary from .52 to .63) and spatial reasoning. It is the quality of the questions that predicts device comprehension, not the quantity. The quality of questions predicted device comprehension better than quality of ideas in a Write Aloud task, although such a difference was not significant. Moreover, the pattern of correlations in the second column in Table 2 (the column for question quality) is extremely similar to the first column (the column for device comprehension scores). In contrast, the correlations in columns 3, 4, and 5 are substantially different from column 1. These results support the major conclusion that the quality of questions asked in the context of a breakdown scenario is a quick litmus test of deep comprehension. Question quality has criterion validity.

There was a modest correlation between the quality of questions and ideas, and also between the volume of questions and ideas. In contrast, there is a modest negative correlation between these volume measures and the two quality measures. So those participants who generate more content tend to produce a lower percentage of quality content. There are several potential interpretations of this unexpected result. Perhaps some students were extremely compliant in producing a large amount of content, even after the high quality content is tapped out. Alternatively, perhaps deep comprehenders are more succinct and discriminating. The available data cannot discriminate between these alternatives.

The fact that the technical knowledge and spatial reasoning measures were robust predictors of device comprehension is quite expected and indirectly confirms the construct validity of the device comprehension measure. Thus, it is the technical knowledge that ends up being important, rather than a host of other measures, such as verbal comprehension and processing speed. Spatial and quantitative components apparently are also important components. However, it should be noted that some of these measures are inter-correlated, so additional analyses are needed to tease apart the contributions of these processes. Openness was the only significant personality measure, perhaps attributable to the creativity component that is linked to this measure. The gender correlations indicate that males have deeper comprehension of devices than females; most researchers would attribute this result to the gender stereotypes in the United States.

Table 2. Correlations of measures collected in the question asking experiment

MEASURES	BIVARIATE CORRELATION COEFFICIENTS				
	Device Comprehension Score	Question Asking		Write Aloud	
		Quality	Volume	Quality	Volume
Cognitive Measures					
ASVAB (g)	.59	.41	.05	.18	.12
Mechanical Reasoning (MR)	.63	.56	-.11	.30	.09
Electronics (EL)	.56	.52	-.01	.36	-.02
General Science (GS)	.60	.48	-.10	.24	.05
Auto & Shop (AS)	.52	.40	-.05	.32	-.05
Mathematics Knowledge (MK)	.56	.45	-.04	.18	.12
Arithmetic Reasoning (AR)	.52	.37	-.05	.11	.04
Numerical Operations (NO)	.12	.10	.13	.10	-.05
Word Knowledge (WK)	.42	.29	.00	.09	.10
Paragraph Comprehension (PK)	.27	.11	-.02	.21	.08
Coding Speed (CS)	.08	-.08	.21	.07	.02
Technical knowledge composite	.72	.55	-.08	.43	-.02
Verbal composite	.49	.28	-.01	.26	.10
Mathematical composite	.59	.44	-.02	.23	.08
Coding and speed composite	.09	.03	.20	.06	-.04
Spatial Reasoning (SP)	.54	.44	.00	.22	.26
Working Memory (WM)	.08	.17	.05	-.08	.12
Exposure to print (EP)	.25	.18	-.11	-.04	.10
Personality Measures (NEO)					
Neuroticism (N)	-.08	.04	-.07	-.07	.01
Extroversion (E)	-.01	-.03	.06	-.03	.02
Openness (O)	.31	.21	-.09	.08	.03
Agreeableness (A)	-.03	-.12	-.02	-.04	-.18
Conscientiousness (C)	-.01	-.04	.18	.01	.13
Demographics Measures					
Gender (GEN, female = 1, male =2)	.41	.33	-.28	.30	-.24
Age	.01	.01	-.01	-.08	-.19
Verbal Protocol Measures					
Quality of questions (QQ)	.51				
Volume of questions (VQ)	-.01	-.34			
Quality of ideas (QI)	.39	.35	-.08		
Volume of Ideas (VI)	.08	-.14	.46	-.34	

Multiple regression analyses. We performed some multiple regression analyses in order to dissect the contributions of the various cognitive components discussed above. Five multiple regression analyses were conducted, one for each of the following five dependent measures: device comprehension scores, question quality, volume of questions, idea quality, and volume of ideas. There were nine predictor variables in each of these analyses: the four ASVAB composite variables, spatial reasoning, working memory, exposure to print, openness, and gender. These predictors were included because they either had a significant correlation with one of the five measures or they were a theoretically important cognitive measure. The results of the multiple regression analyses are presented in Table 3.

Table 3. Beta weights in multiple regression analyses in the question asking experiment

MEASURES	DEPENDENT MEASURES				
	Device Comp	Question Asking		Write Aloud	
	Score	Quality	Volume	Quality	Volume
Cognitive Measures					
ASVAB					
Technical knowledge	.51*	.37*	.24	.47*	.15
Verbal	.10	-.16	.01	.19	.00
Quantitative	.14	.13	-.02	-22	.03
Speed	.06	.01	.09	.20*	.17
Spatial Reasoning (SP)	.11	.15	.00	-.01	.39*
Working Memory (WM)	-.05	-.02	.00	-.11	.08
Exposure to print (EP)	-.12	.16	-.17	-.27*	.13
Other Measures					
Openness (O)	-.02	-.06	-.08	-.08	-.03
Gender (GEN, female = 1, male =2)	.11	.13	-.37*	.19	-.29*
Variance Predicted (R^2)	.57*	.37*	.13	.26*	.19*

* Statistically significant at $p < .05$

The multiple regression analysis for the device comprehension scores revealed that technical knowledge was the primary predictor variable. The multiple regression equation with 9 predictors accounted for 57% of the variance; technical knowledge alone accounted for 52% of the variance (i.e., $.72^2 = .52$). None of the other 8 predictors were significant. We also assessed interactions between pairs of predictor variables, but these were rarely significant (less than 5%, readily attributable to a Type 1 error). So it is technical knowledge that reigns supreme in predicting device comprehension scores, an outcome that confirms the construct validity of our gold standard of deep comprehension.

The multiple regression analysis for question quality perfectly mirrored the results of the device comprehension scores. The multiple regression equation with 9 predictors accounted for 37% of the variance. Technical knowledge alone accounted for 26% of the variance (i.e., $.51^2 = .26$) and none of the other 8 predictors were significant. Once again, pairwise interaction components also were rarely significant. This result supports the earlier claim that question quality is an excellent index of deep comprehension. The measure has a satisfactory degree of construct validity.

The remaining three measures of the verbal protocols did not fare as well. The volume of questions was not significantly predicted by the 9 variables. The volume of ideas in the Write Aloud Task was significantly predicted by spatial reasoning and gender, with females articulating more information. The quality of ideas was significantly predicted by technical knowledge, speed and exposure to print (the latter in the opposite direction, a likely suppression effect).

We performed a factor analysis that included the 4 variables that comprised technical knowledge and the remaining 3 cognitive measures (spatial reasoning, working memory span, and exposure to print). The results revealed that there was only one significant underlying factor, with an eigenvalue of 3.76 (chance being 1.38). The four ASVAB measures loaded most heavily on the factor. We interpret this underlying factor to be deep comprehension of technical knowledge. Device comprehension scores and question quality are two excellent measures of this factor.

Qualitative analyses of questions. Technical scientific knowledge was highly correlated with device comprehension scores so we examined the differences between the questions that were asked by participants with high versus low technical knowledge. In one analysis, we compared the distributions of questions that were asked by college students with high scores (upper 33% of distribution) versus low scores (lower 33% of the distribution). The questions asked by students with high scores had two characteristics: (a) the questions converged on components in the mechanism that are plausible faults and (b) the questions had a more fine-grained elaboration of the parts, processes, and relations that specify how the breakdown occurred. Stated differently, there was high convergence and high mechanistic detail. In contrast, the questions asked by students with low scores were quite different: (a) the questions were diffuse rather than convergent (e.g., most of the components in the system were broken instead of converging on 1 or 2 components), (b) the questions had minimal elaboration of the processing mechanism (e.g., component X was broken but there was no elaboration of how it might be broken), and (c) some questions were shallow or irrelevant.

In order convey some of the above observations more concretely, Table 4 presents the distribution of questions asked for the cylinder lock. This table was reproduced from an article by Graesser, Olde, Pomeroy, et al. (in press). These include all questions that were generated by at least 2 participants with high mechanical comprehension scores (out of 11 participants) or with low mechanical comprehension scores (out of 11); only half of the sample of participants were available at the time of this analysis. Three points can be gleaned from these data. First, students with high mechanical comprehension scores identified the fault in the region of the cam with a much higher incidence than did the low mechanical students. Second, the amount of mechanical detail is quite striking in the case of participants with high mechanical comprehension, and quite underwhelming in the case of those low mechanical comprehension. These first two observations support the claim that deep comprehenders ask better questions. Third, the volume of questions was the same for the two groups of participants. So deep comprehenders do not simply ask more questions; rather they ask questions of higher quality. Fourth, agreement was much higher for those with high mechanical comprehension. Agreement is measured as the proportion of questions that were generated by 2 or more participants within an ability group (as opposed to idiosyncratic questions generated by only one participant). Therefore, deep comprehenders converge on likely malfunctions in the device whereas shallow comprehenders sample ideas haphazardly and indiscriminately.

In one other analysis we examined the nodes in the conceptual graph structures and segregated them into the following five categories.

Fault with large part. The node was a fault of the breakdown and had at least one large part.
Fault with small part. The node was a fault of the breakdown, but had no large part.

Nonfault with large part. The node was not a fault of the breakdown, but had at least one large part.

Nonfault with small part. The node was not a fault of the breakdown and had no large part.

Other. The node was a goal or had no parts.

For each of the nodes in the first four categories, we computed the likelihood that a participant articulated a question about the node. The likelihood values were .46 and .27 for the fault nodes with big versus little parts, respectively. So clearly, the size of the components had an influence on whether a node was articulated. The likelihood values were .19 versus .10 for the nonfault nodes with big versus little parts, respectively.

Table 4: Questions About A Cylinder Lock.

	MECHANICAL COMPREHENSION	
	HIGH	LOW
Is it the right key?	6	5
What kind of lock is it?	0	3
Is the spring broken?	5	5
Does spring keep bolt from moving?	3	0
Is the spring pulling back the bolt?	2	0
Is the spring making the bolt get stuck?	2	0
Is the cam broken?	6	2
Is the cam moving?	2	2
Is the cam moving back the bolt?	2	0
Is the bar that fits under cam broken?	2	0
Is the cam disconnected/out-of-synch with the cylinder?	2	0
Is the cylinder turning?	3	2
Is the cylinder turning the cam?	2	0
Is the bolt stuck in the slot?	3	3
Is the bolt connected to the bar?	2	0
Are the pins broken?	4	2
Do the pins lift right?	3	2
What are the pins used for?	0	2
Number if participants	11	11
Questions per participant	7.0	7.1
Proportion of common questions	.70	.35

4. Eye Tracking and Question Asking: An Empirical Study

Eye tracking provides an important window for dissecting the cognitive processes and representations that play a role in particular cognitive tasks. Examples of such tasks are the comprehension of sentences (Rayner & Polletsek, 1989), the comprehension of text (Just & Carpenter, 1980; Magliano, Graesser, Eymard, Haberlandt, & Gholson, 1993; O'Brien, Raney, Albrecht, & Rayner, 1997), the comprehension of illustrated text (Hegarty & Just, 1993), the perception of real world scenes (Loftus & Mackworth, 1978), the evaluation of arguments (Wiley, 1999), and reasoning (Just & Carpenter, 1992). However, rigorous eye tracking research is conspicuously lacking in several arenas that are relevant to complex information and communication technologies, such as question asking, question answering, hypertext, graphic displays, animation, and computer simulations.

At this point, no one has systematically analyzed the relationships between eye tracking and the cognitive components in question asking. There are a number of hypotheses that could be directly derived from the PREG model. We would expect a high density of eye fixations to occur at words, objects, parts, and processes that are at the source of cognitive disequilibrium (e.g., anomalies, contradictions, broken parts, contrasts, missing components, and so on). However, it should take a sufficient amount of technical knowledge to detect such irregularities in the system. There should be a systematic relationship between eye tracking behavior and technical knowledge when college students generate questions in the context of a breakdown scenario. Students with high technical knowledge should focus on the causes of the device breakdown (e.g., the cam) whereas students with low technical knowledge should indiscriminately scan the regions of the illustrated text. That is, technical knowledge and other indices of deep comprehension should be positively correlated with the percentage of fixations and the percentage of time that the comprehender focuses on the fault area. These predictions were tested in the empirical study reported in this section.

Software has been developed to provide area plots and gaze traces after eye tracking data have been collected on a display (Marshall, 1999). An area plot displays the amount of time that the eye fixates at each region in an N x M dimensional grid. The area of interest is the subset of the display that should theoretically receive fixations (i.e., the faults of a malfunctioning device). The area plot is to be contrasted with a gaze trace, which plots the sequence of eye fixations at X-Y coordinates as a function of time. When examining the gaze traces, we would expect eye movements to drift toward a locus of disequilibrium (fault) immediately before or during the articulation of a question. The present study investigated the patterns of eye tracking that occurred before, during, and after the articulation of a question.

Methods

Participants. The participants were 40 college students at the University of Memphis. The students participated for course credit in an introductory psychology class.

Illustrated texts and tasks. The participants read 5 illustrated texts on everyday devices: a cylinder lock, an electronic bell, a car temperature gauge, a toaster, and a dishwasher. These were the same devices that were used in the previous study on question asking. The clutch was dropped from the analysis because it was extremely difficult for participants to differentiate and label the individual teeth in the wheels of the clutch mechanism. As in the previous study, each of the five trials consisted of two phases. The participant first read the illustrated text for 3 minutes, which was displayed on a computer monitor. After the reading phase, the breakdown description was presented either above or to the left of the illustrated text and the participant began the question asking phase (while the illustrated text remained on

the screen). The participants asked questions aloud for 90 seconds during this phase and the protocol was recorded. The previous study had participants generate questions in writing whereas the present study collected spoken questions. Each participant furnished question asking protocols for all 5 devices. The assignment of devices to test order was counterbalanced across the 40 participants with a Latin square.

The participants completed a number of tests after they read the illustrated texts and generated questions. They answered the 30 questions on the device comprehension test (5 devices X 6 three-alternative, forced-choice question per device). They subsequently completed assessments of the following measures of individual differences: ASVAB's four scales of technical knowledge (mechanical comprehension, electronics, general Science, auto & shop), spatial reasoning, and openness. These were the statistically significant predictors of deep comprehension and question asking in the previous study.

Recording of eye tracking and question asking. Eye movements were recorded by a Model 501 Applied Science Laboratory eye tracker. There was a head mounted recording unit so the participants could move the head during data collection. The participants were calibrated before they started the experimental session of reading the illustrated texts and asking questions. During calibration, the participants viewed 9 points on the computer display and a computer recorded the x-y coordinates. The equipment, computer, and focus of the eye gaze became synchronized after these recordings. The calibration process took 10-15 minutes, depending on the pupil size and other parameters. Participants were dismissed if they wore glasses, but the equipment could accommodate contact lenses.

The experimental session was videotaped and audio recorded. The camera focused on the computer screen. The VCR recorded the illustrated text displayed on the screen and a superimposed image of what the left eye was focusing on. The superimposed image was generated by the eye tracking equipment. The superimposed image showed the locus of (a) the focus of the eye and (b) an X-Y axis with the 0-0 point at the center of the focus. The voice of the participant was recorded on the VCR so that the spoken questions could be transcribed. This set-up allowed us to record and review (a) the contents of the computer display, (b) the focus of the left eye, and (c) the voice of the student asking questions.

Computer software was available to record eye tracking behavior at a fine-grained level. The software produces area plots for specific areas of interest. In particular, we were interested in the percentage of time and the percentage of eye fixations in the areas of interest associated with faults. These faults were sometimes in the text and sometimes in the picture. We were interested in a gaze trace before, during, and after the articulation of the question. During these time spans, the software printed out a sequence of numbers at locations associated with the eye focus.

Results and Discussion

Descriptive statistics. Table 5 presents descriptive statistics on the measures collected in the eye tracking experiment. The means and standard deviations of the measures of individual differences were very similar in this experiment and the question asking study reported in the previous section (see Table 1). Similarly, the volume of questions and quality of questions were comparable. The device comprehension score was 18.6 for the five devices in this experiment, which is 62% of the questions being answered correctly. This is comparable to the 65% in the previous experiment that had 6 devices tested. Regarding the eye tracking measures, there were 29.5 fixations on plausible faults per device, or 9.3 seconds out of 90 seconds. The percentage of eye fixations that were on faults was 11.5%, whereas the percentage of time on the faults was 10.4%. It should be noted that the percentage of time on faults and the total fault fixation time are functionally equivalent because the participant was always allocated 90 seconds per device for question asking.

Table 5. Descriptive statistics on measures collected in the eye tracking experiment

MEASURES	DESCRIPTIVE STATISTICS Mean	Standard Deviation
ASVAB		
Mechanical Reasoning (MR)	13.9	6.0
Electronics (EL)	9.3	4.3
General Science (GS)	16.9	4.7
Auto & Shop (AS)	10.0	5.3
Spatial Reasoning (SP)	23.8	15.2
Openness (O)	52.1	9.8
Gender (GEN, female = 1, male =2)	1.25	.44
Verbal Protocol Measures		
Volume of questions (VQ)	5.3	2.9
Quality of questions (QQ)	29.3%	16.5
Eye Tracking Measures		
Number of fault fixations per device	29.5	11.1
Percentage of fixations on faults	11.5%	4.1
Total fault fixation time per device	9.3	3.9
Percentage of time on faults	10.4%	4.3
Device Comprehension Score	18.6	4.6
Number of Participants	40	

Correlations. Table 6 presents bivariate correlations between measures in the eye tracking experiment. As in the previous study, the device comprehension score was regarded as the gold standard of deep comprehension. This experiment replicated the previous experiment in showing that device comprehension scores were significantly correlated with the ASVAB technical knowledge composite measure (and all of its component measures), spatial reasoning, openness, and gender. The number of questions asked about plausible faults also significantly correlated with device comprehension scores, whereas the total volume of questions did not. Moreover, all three measures of the eye tracking performance significantly correlated with device comprehension scores: the number of eye fixations on faults, the percentage of fixations on faults, and the total time fixating on faults. Thus, eye tracking has criterion validity in predicting deep comprehension. A valid litmus test of deep comprehension is whether the participant spends a greater percentage of time focusing on plausible faults when faced with a breakdown scenario.

The magnitude of the correlations support the claim that fixating on faults is a robust indicator of deep comprehension. For example, the ASVAB technical knowledge composite score had a .54 correlation with device comprehension scores; the proportion of time the eye fixated on faults had a .50 correlation with device comprehension scores. So a 90 second clip of eye tracking data was just as valid as a 2-hour paper and pencil test that has survived multiple standards of psychometrics. One of the advantages of the eye tracking data is that deep comprehension can be assessed for specific devices, whereas the scope of an ASVAB test is generic rather than specific.

Table 6. Correlations in eye tracking experiment

MEASURES	DEPENDENT MEASURES			
	Device Comp Score	Focus on Faults		
		Number of fixations	Percentage of fixations	Time on faults
ASVAB Technical knowledge	.54*	.31*	.28	.33*
Mechanical Comprehension	.39*	.16	.14	.15
Electronics	.32*	.18	.23	.19
General Science	.60*	.49*	.43*	.51*
Auto & Shop	.58*	.25	.18	.29
Spatial Reasoning (SP)	.45*	.15	.16	.17
Openness (O)	.41*	.31*	.36*	.41*
Gender (GEN, female = 1, male =2)	.45*	.27	.18	.27
Device Comprehension Score	--	.43*	.31*	.50*
Volume of Questions	.20	.08	.20	.21
Number of Fault Questions	.45*	.42*	.49*	.52*
Quality of Questions	.24	.26	.24	.24

* Statistically significant at $p < .05$

Eye tracking when questions are asked. We conducted follow-up analyses that focused on the good questions that were asked. We were curious about the coordination of eye fixations with the asking of good questions, namely those questions that focused on plausible faults. We computed the percentage of eye fixations in a fault region as a function of (a) high versus low technical knowledge, as measured by ASVAB, and (b) time slices (3 seconds before the question, during the question, versus 3 seconds after the question). A median split criterion was used to segregate participants into high versus low technical knowledge. High technical knowledge participants had percentage scores that were 14.3%, 11.9, and 10.6 for before, during, and after the question, respectively. The corresponding percentages for participants with low technical knowledge were 9.0%, 9.6, and 7.2, respectively. An analysis of variance was performed on these percentages, using a Knowledge x Device x Time-slice design. Knowledge was a between-subjects variable whereas Device and Time-slice were within-subjects. There was a statistically significant main effect of knowledge, $F(1, 38) = 4.75$, $p < .05$, $MS_e = 6.0$, device, $F(4, 152) = 15.86$, $p < .05$, $MS_e = 4.3$, and time-slice, $F(2, 76) = 3.41$, $p < .05$, $MS_e = 1.1$, but no significant interactions. Regarding the time-slide, there was a gradual decrease in percentages as one moved from before, to during, to after the questions, 11.7, 10.8, versus 8.9, respectively. These results suggest that participants often look at the faults before the questions are launched.

We performed an analysis on the qualitative patterns of eye tracking that occurred while questions were asked about plausible faults. We isolated those questions that tapped plausible faults and observed the VCR film clips in the stretch of time between 3 seconds prior to the launching of the question to 3 seconds after the completion of the question (about 9 seconds on the average). Trained judges observed the films and classified the sequence of eye fixations into one of the following seven categories.

(1) Causal process flow
(2) Focus on causal antecedents of fault
(3) Focus on causal consequences of fault
(4) Integration of text and picture
(5) Integration of text and breakdown scenario description
(6) Back and forth between two picture components
(7) Focus on pictures

It should be noted that a single 9-second observation could be classified into more than one of these categories. The percentage of observations in these seven categories was 23%, 21, 26, 55, 6, 33, and 19, respectively. An ANOVA was performed on these categories, using a technical knowledge (high versus low) by category factorial design. There was no significant main effect of knowledge and no significant interaction, but the main effect was significant for category.

Conclusion

In closing it appears that we have two quick tests of whether adults have deep knowledge about a particular device. In both tests, we present a breakdown scenario that puts the participants in cognitive disequilibrium and that forces a problem solving mode. One test is that they will generate good questions that tap likely causes of the breakdown. The second test is that their eyes tend to fixate on the faults. In contrast, the poor comprehenders have questions that are not discriminating and their eyes move more indiscriminately over the display. In less than 2 minutes, we can identify whether a particular sailor has the deep knowledge and talent for understanding a particular device.

5. PREG: A Model of Question Asking

PREG is a model of human question asking (Graesser, Olde, Pomeroy, et al., in press; Otero & Graesser, in press). The model contains a set of production rules that specify the conditions under which adults ask questions when they read expository texts. The essence of PREG's question asking mechanism is the existence of discrepancies between the representation of text information and the reader's world knowledge, with a mediating role of pragmatics and metacognition. Both the explicit text and the world knowledge are represented in the form of a conceptual graph structure. Comparisons between text representations and readers' knowledge are carried out by examining the three components of conceptual graph structures: words, statements, and links between statements. The predictions of PREG are presented in this section. Support for these predictions are reported in Otero and Graesser (in press).

Background Research on Question Asking

Question asking has frequently been considered a fundamental cognitive process in the field of education (Dillon, 1988; Fishbein, Eckart, Lauer, van Leeuwen, & Langmeyer, 1990; Flammer, 1981; King, 1989, 1992, 1994; van der Meij, 1988; Zimmerman, 1989). The ideal learner is an active, self-motivated, creative, inquisitive person who asks deep questions and searches for answers to such thought-provoking questions. There is a long history of researchers who have advocated learning environments that support inquiry learning and the acquisition of self-regulated learning strategies (Bransford, Goldman, & Vye, 1991; Collins, 1988; Piaget, 1952; Palincsar & Brown, 1984; Papert, 1980; Pressley & Levin, 1983).

The disappointing news is that most students are not vigorous question askers, and most educational settings do not support student question asking. For example, it is well documented that students rarely ask questions in classrooms and most of their questions are shallow (Dillon, 1988; van der Meij, 1988; Graesser & Person, 1994). Graesser and Person (1994) reported that an average student asks only .1 question per hour in a classroom; this rate of question asking substantially increases in one-to-one human

tutoring (26.5 questions per hour), but the vast majority of these questions are shallow questions rather than questions that promote deep reasoning (e.g., why, how, what-if, what-if-not). However, there are reasons to be optimistic about the prospects of developing learning environments that improve question asking and learning. There is ample empirical evidence that students can be trained to ask good questions and that such training leads to significant gains in learning and literacy (Beck, McKeown, Hamilton, & Kucan, 1997; Davey & McBride, 1986; King, 1989, 1992, 1994; Palincsar & Brown, 1984; Singer & Donlan, 1982). We believe that a sophisticated understanding of question asking should strengthen this link between question asking and learning.

Existing research on question asking has uniformly embraced the notion that clashes between stimulus input and world knowledge are very much at the essence of question generation (see Introduction). Thus, questions are asked when there are contradictions, anomalous information, obstacles to goals, uncertainty, and obvious gaps in knowledge. Although it is widely acknowledged that discrepancies between input and knowledge trigger questions, the precise mechanisms need to be specified in more detail than has been achieved in psychology and education.

The field of artificial intelligence has offered computational models that make some attempt to specify the knowledge representations and knowledge discrepancies that underly question asking (Kass, 1992; Reisbeck, 1988; Schank, 1986, 1999). According to Schank's (1986) SWALE model, for example, questions are asked when we observe anomalous events and ask questions that explain such events (such as "Why did the event occur?"). Long-term memory is viewed as a large inventory of cases that record anomalous events and their associated explanations (which are driven by why, what-if, and other deep questions). Unfortunately, these models in artificial intelligence have never been tested by collecting data on humans, so we are uncertain about the extent to which these models mirror human cognition. The present research was expected to reduce the large gap that exists between the precise computational models in artificial intelligence and the empirical research in education and psychology.

The PREG model predicts the particular questions that adults ask when they read expository texts on scientific phenomena. The predicted questions are sensitive to four information sources or processing components: (1) the explicit text, (2) the reader's world knowledge about the topics in the text, (3) the reader's metacognitive skills, and (4) the reader's knowledge about the pragmatics of communication. Although the complete PREG model is sensitive to the reader's metacognitive skills and knowledge of pragmatics, this section concentrates on the process of generating questions on the basis of the explicit text and the reader's world knowledge. Metacognition and pragmatics will be addressed, as needed, when they offer illuminating predictions. The PREG model contains a set of production rules, that identify the particular conditions that produce particular questions. The questions are sensitive to features of the explicit text and world knowledge.

The PREG model adopts a theory of knowledge representation and a production rule formalism. Both the explicit text and the world knowledge are represented as conceptual graph structures (Graesser & Clark, 1985; Graesser et al., 1992). These structures map out the causal chains, goal hierarchies, taxonomic hierarchies, spatial composition, and properties of the domain knowledge under consideration (see Figure 2). A production rule is an "IF <condition> THEN <action>" formalism which specifies the particular cognitive or behavioral actions that are activated when particular conditions exist in the system (Anderson, 1983; Just & Carpenter, 1992; VanLehn, 1990). The conceptual graph structures and production rules together provide a suffient level of analytical detail to capture the systematic mechanism of question asking.

Discrepancies as the Basis for Question Asking

The essence of PREG's question asking mechanism is the existence of discrepancies between the representation of text and the reader's domain knowledge about the topics in the text. However, there is a nontrivial relationship between text and world knowledge as triggers for questions. This can be illustrated in the following two hypotheses that make quite different predictions.

Many questions about the text will not be asked if a reader lacks the appropriate knowledge to be compared with the representation of the explicit text. A knowledge clash hypothesis predicts more questions as a function of increasing world knowledge because there is a greater incidence of incompatibilities between the text and world knowledge. A similar prediction is made by Miyake and Norman (1979) who argue that it takes a large amount of knowledge to know what one does not know. A simple knowledge deficit hypothesis would make quite different predictions. It predicts that the number of questions should decrease as a function of increasing world knowledge because there is less uncertainty and fewer gaps in knowledge. This knowledge deficit hypothesis is consistent with the fact that readers sometimes ask questions when they do not have knowledge that clashes with the text. For example, when readers encounter a rare word, such as *cam*, they frequently ask what the word means (*What does cam mean?*). We believe that both of these hypotheses have some validity. Questions are triggered by discrepancies in both cases: the difference lies in the nature of the representation in the text and in the knowledge of the reader. The impact of discrepancies on questions can be unpacked further by dissecting the different levels of text representation.

Comparisons between text representations and readers' knowledge can be made for the three components of conceptual graph structures: words, statements, and links between statements. The PREG model examines discrepancies for these three components. First, word-triggered questions occur when there are words with unknown meaning, or words with unknown or ambiguous referents. The simplest case is a question on a completely unknown word. In the comprehension monitoring literature, this is an application of the "lexical standard" for comprehension monitoring (Baker, 1985). Second, statement-triggered questions are asked when readers are unable to adequately represent a statement (i.e., state, event, goal) in the textbase or situation model. Simply put, the reader has trouble constructing the meaning for an explicit statement in the text. There are many reasons for the failure to construct a meaning. The reader is unable to either (a) create a mental model that meshes with the statement in the textbase, (b) relate the textbase statement to an existing representation in the mental model, or (c) resolve a discrepancy between the textbase statement and the readers' background knowledge. Baker's (1985) "external consistency" standard is adopted when a reader notices this last type of discrepancy in "c". Finally, link-triggered questions are caused by an inability to represent a link at the mental model level, or by a mismatch between an explicit text link and the reader's knowledge about the appropriate link. This section describes and explains the discrepancies that exist for the three constituents (word, sentence, versus link) and the different levels of text representation.

Word level. A word-triggered question is generated when a reader is uncertain about the meaning of a particular word in the text. This may happen because the word is completely unknown to the reader or because no referent is found for it, even when the meaning is known. Thus, there may be a discrepancy between a word in the text and (a) the lexicon of word knowledge or (b) the referent of the word in the mental model. These word level questions are the most frequently asked questions in most learning enviroments (Graesser & Person, 1994).

(A) Unknown word. A reader may be ignorant of the meaning of a word.

IF A content word X (noun, main verb, or adjective) in the text is not known
THEN Ask: "What does X mean?"

(B) Unknown referent. The explicit text mentions a noun or pronoun, but it is difficult to construct or identify a referent in the mental model that corresponds to the noun/pronoun.

IF Referent of a noun or pronoun X is not known
THEN Ask: "What X?"

(C) Ambiguous referent. A noun or pronoun in the text can refer to more than one referent in the mental model.

IF Referent of a noun or pronoun X can refer to more than one referent
THEN Ask: "Which X?"

Statement level. Questions at the statement level directly depend on the reader's world knowledge. This knowledge may be stored in episodic representations, or in semantic representations that have been formalized as schemata and generic knowledge structures (Graesser & Clark, 1985; Lenat, 1995). Statement-triggered questions may have two origins: incomprehensible statements or discrepant statements. In the first case, a reader is unable to create a referent or mental model representation for a statement in the textbase. In the second case, there is clash between reader's world knowledge and the representation of a text statement at the mental model level.

(D) Incomprehensible statements. The reader is unable to create a referential representation of the information explicitly stated in the text. In order to solve this problem, readers may ask directly "What does statement S mean?" or formulate the inquiry as a "How" question.

IF Statement X can not be represented at the mental model level
THEN Ask: "What does X mean?" or "How does X occur/exist?"

(E) Discrepant Statement. Questions are asked when an explicit statement in the text is discrepant with a reader's knowledge of the explicit text or with implicit knowledge. A clash between an explicit text statement and prior explicit text is easiest to detect and specify theoretically. Clashes with implicit knowledge are more subtle, particularly when the central foundation lies in metacognition and in opaque features of language and discourse.

IF Statement X clashes with world knowledge (see E1 through E5)
AND No incoming Consequence or Implies link feeds into X in the textbase
THEN Ask: "Why did X occur/exist?" or "How did X occur/exist?"

E1 to E5 below are are attempts to tune the precise conditions in which discrepancies are detected and trigger questions.

(E1) Inconsistencies. The reader has world knowledge that clashes with the text statement.

(E2) Given versus new information. The PREG model does not assume that all discrepancies with world knowledge are queried. If that were the case, there would be massive questioning by students. According to the PREG model, readers are more proned to ask questions about "new" information than "given" information in the text. The writer assumes that the reader already knows and accepts the given information, and is informing the reader about new content (Clark, 1996; Haviland & Clark, 1974). The given information is presupposed to be true, whereas the new information is potentially under scrutiny. The Moses illusion is an excellent example of our tendency to gloss over and automatically accept

presupposed information as being true (Reder & Cleeremans, 1990). When asked "How many animals of each kind did Moses have on the ark?", most people quickly say "two" instead of pointing out that it was Noah, not Moses, who had animals on the ark. The discepancy gets missed because the question presupposes, rather than asserts, that Moses had animals on the ark. When asked "Was it Moses who had animals on the ark?", most people quickly say "No, it was Noah who had animals on the ark." The PREG model assumes that there is a higher likelihood of questions being triggered by new information than presupposed, given information.

(E3) **Initial steps in causal chains**. The reader tries to explain why states exist and events occur. Answers to such why-questions trace the causal antecedents that lead up to the events (Graesser & Hemphill, 1991). According to some theories of comprehension (Graesser, Singer, & Trabasso, 1994; Schank, 1999), readers attempt to formulate explanations whenever they comprehend virtually any type of text. That is, readers attempt to explain why actions, events and states are mentioned in a text, and why they exist or occur in the mental model. Attempts to achieve explanations are particularly prevalent when reading expository texts on devices and scientific mechanisms (Bertus & Graesser, 1998; Millis & Graesser, 1994; Singer, 1994). Millis and Graesser (1994) and Graesser and Bertus (1998) reported that readers of scientific texts generate more causal antecedent inferences than consequence inferences (i.e., expectations about future events) and other types of elaborative inferences. Given that readers have a strong tendency to search for causal antecedents of states and events, they would be expected to ask causal questions when the text or their world knowledge fails to find a cause of an explicit statement in the text. It follows that many why-questions should be asked about the first step in a causal chain that is expressed in a text. Suppose that a text explicitly articulates a causal chain of events: E_1, E_2, ... E_n. There should be many questions about event E_1 because the text does not elaborate the causal antecedents that explain why or how it occurred. There should be few, if any, questions about event E_n because it is explained by the explicit chain of causal antecedent events E_1 through E_{n-1}. Therefore, PREG predicts that the first steps in explicit causal chains should trigger why-questions more often than subsequent steps in a causal chain.

(E4) **Negations**. Writers do not have the habit of constructing expressions with negation (i.e., not P), at least not capriciously. It would be possible to generate thousands of negative expressions about any mental model (e.g., a cam is not a cylinder, a cam is not square, the cylinder does not rotate). Writers do not articulate a massive amount of negative expressions. In fact, expressions with negation account for less than 5% of statements in a textbase and in verbal protocols produced by college students in experiments (Graesser & Clark, 1985). Writers save negative expressions for situations when a comprehender might believe P or hope that P is true, but somewhat expectedly not-P is the true state of affairs. In a sense, the writer is implictly using the following rhetorical frame: "You (the reader) might believe, want, or hope that P is true, but actually not-P is true." Therefore, readers implicitly ask "Why not?" whenever a negation appears in the text. If an answer to the Why-not questions fails to exist in the textbase or world knowledge base, then the question will be asked overtly. Therefore, PREG predicts that negative expressions in a text will have a higher likelihood of being flagged with why-questions than do positive expressions.

(E5) **Extremely precise content**. Extremely precise content in the text is a magnet for questions when the surrounding content is imprecise. Why? Because the extremely precise expressions violate the Gricean maxim of quantity. According to this maxim, a cooperative writer should not be more specific than is required in the communicative context. If the vast majority of the events in a text are not embellished with precise time specifications, but all of a sudden one event does have a precise time index (e.g., in August, at 2:00 am), then there is a discrepancy between the prevailing style of the text and the event with a precise time index. The reader will implicitly ask the question "Why is the writer being so precise about the time for this event?". If the reader cannot construct a reason, the question will be asked overtly.

Link level. Link-triggered questions result from a discrepancy between activated world knowledge structures and links existing in the text. These links are sometimes signaled by explicit connectives (*so, because, consequently, in order to, so that*). Sometimes a causal link is signaled by the contiguity of cause and effect in the surface structure of a text, without a direct specification of the relation. The links that the PREG model adopted are the relations in the conceptual graph structures (Graesser et al., 1992, see Figure 2). The world knowledge that is especially relevant to expository texts on science topics includes the following relations: Consequence (e.g., cause or enable), Implies, Manner, Property and Set Membership. The Consequence link designates a causal relation between two events. It is directed, such that the first event/state precedes the resulting event/state in time. The Implies link is similar to the Consequence link, except for the temporal constraints that exist between the source node and the end node. For the Consequence link, the source node precedes the end node in time, both nodes exist simultaneously for the Implies link. Manner links specify the speed, style, or other dynamic characteristics of an event; an event node is connected to another event node that elaborates its style, and the two events unfold simultaneously in time. The Property links elaborate the properties of a concept or a concept embedded in a statement node. Finally Set Membership links correspond to class inclusion relation (is-a).

(F1) **Incomprehensible Consequence or Implies Link**.

> IF Consequence or Implies link L connecting statements X and Y is not comprehensible
> THEN Ask: "Why Y", "How X L Y?"

(F2) **Incomprehensible Manner Link**

> IF Manner link L connecting statements X and Y is not comprehensible
> THEN Ask: "How X L Y?"

Similar production rules can be formulated for Property and Set Membership links.

(G1) **Discrepant Consequence or Implies Link**

> IF Consequence or Implies Link L connecting statements X and Y clashes with world knowledge
> THEN Ask: "Why Y" or "How X L Y?"

(G2) **Discrepant Manner Link**

> IF Manner Link L connecting statements X and Y clashes with world knowledge
> THEN Ask: "How X?" or "How X L Y?"

Similar production rules can be formulated for Property and Set Membership links.

Tests of Predictions of PREG Model

Otero and Graesser (in press) tested the predictions of the PREG model by having participants generate questions while they read expository texts on scientific mechanisms. It is beyond the scope of this final report to discuss these analyses. However, we will point out two findings that are noteworthy. First, the PREG model could account for over 90% of the questions that students asked about the scientific texts. Therefore, PREG had a high recall score, using the standard terminology in computational linguistics. Second, the PREG model had a respectable precision score; this is the proportion of theoretically

predicted questions by PREG that participants actually asked. The combination of recall and precision scores supported the claim that PREG was quite discriminating in predicting when questions are versus are not asked while readers comprehend scientific text.

6. Practical Implications

This research is most directly relevant to the selection and classification of personnel in the Navy and to training. However, the research is also likely to benefit many other Navy missions that were articulated in ***Sailor 21: A Research Vision to Attract, Retain, and Utilize the 21st Century Sailor.*** Listed below are some of the salient applications.

(1) **Assignment of personnel to jobs that use equipment**. The process of personnel selection and assignment involves appropriate matches between the abilities of personnel and the jobs assigned to them. ASVAB is currently used as a psychometric test that measures important cognitive components, namely general intelligence, verbal, numerical, technical, and speed. These tests have demonstrated some degree of reliability and validity, but improvements can be made to the extent that they are grounded in research in cognitive science. Our research has revealed that ASVAB's mechanical, electronic, and general science subtests do an excellent job predicting the depth of a person's comprehension about devices in general. However, these tests have two drawbacks: (a) they take a long time to administer (2-3 hours) and (b) they do not directly predict performance on a particular device.

The present research has demonstrated much quicker, device-specific assessments of deep comprehension of devices. An adult is given a breakdown scenario and generates questions while the eye tracking equipment records the eye fixations and eye movements. Deep comprehenders generate better questions that converge on the likely faults of the breakdown, whereas shallow comprehenders ask less discriminating questions. The eye movements of deep comprehenders focus on the likely faults, whereas those of the shallow comprehenders are less discriminating. In less than 1 or 2 minutes, a valid assessment can be made about the depth of a person's understanding of a particular device. A quick assessment can be critical in a wartime situation where sailors need to rotate the use of particular equipment under extreme time constraints.

(2) **Better assessments of deep comprehension of devices**. For many jobs, personnel need to be selected and trained on the basis of how well they can operate, repair, and maintain particular devices. Deep comprehension is necessary when the devices malfunction; shallow knowledge and jargon do not go the distance. Once again, however, the available psychometric tests on reasoning tap general abilities rather than the knowledge and ability to reason about particular devices. Theoretical guidance is needed on how to design a test on a specific device. That is, given that a device is introduced to personnel in the military, how can researchers quickly design a test that assesses the personnel on the device knowledge? The present project has validated three methods of assessing deep comprehension about a device: (a) question asking in the context of a breakdown scenario, (b) eye tracking in the context of a breakdown scenario, and (c) a 3-alternative, forced-choice test under the guidance of theories of qualitative physics.

(3) **Design of query modules in computer-human interfaces**. The design of most information and communication technologies requires some theoretical guidance in accommodating user questions and answers to such questions. What questions do users have? How should the questions be answered? PREG provides a theoretical foundation for researchers in human-computer action who want to design an effective conversational interface.

(4) **Frequently Asked Questions (FAQ's)**. FAQ's are a popular facility in most computer applications. The PREG model of question asking provides the foundation for generating a list of likely questions in a FAQ module. Most questions in a FAQ facility are generated by the software designers rather than a sample of end users. The production rules of PREG will provide some guidance for the designer in generating the sample of likely questions.

(5) **Question-answer modules in hypertext and hypermedia**. These systems on the Web and CD-ROM are supposed to handle questions that users have while exploring some domain of knowledge. PREG provides some guidance in handling the space of questions that users will have about particular content.

(6) **Deep learning**. Learners need to be conceptually challenged with difficult problems in order to gain deep comprehension of a complex mechanism. Breakdown scenarios present a suitable challenge because breakdowns frequently occur in the real world and there is a pressing need to fix such breakdowns. Learning environments that are built around equipment breakdowns are motivating, have high ecological validity, and are pedagogically effective.

7. References

Anderson, J.R. (1983). The architecture of cognition. Cambridge, MA: Harvard University Press.

Baggett, W.B., & Graesser, A.C. (1995). Question answering in the context of illustrated expository text. Proceedings of the 17th Annual Conference of the Cognitive Science Society (pp. 334-339). Hillsdale, NJ: Lawrence Erlbaum.

Baker, L. (1985). How do we know when we don't understand? Standards for evaluating text comprehension. In D.L. Forrest-Pressley, G.E. Mackinnon, T.G. Waller (Eds) Metacognition, cognition and human performance (pp. 155-205). New York: Academic Press.

Beck, I.L., McKeown, M.G., Hamilton, R.L., & Kucan, L. (1997). Questioning the Author: An approach for enhancing student engagement with text. Delaware: International Reading Association.

Bennet, G.K., Seashore, H.G., & Wesman, A.G. (1972). Differential aptitude test: Spatial relations., Form T. New York: Psychological Corporation.

Berlyne, D.E. (1960). Conflict, arousal, and curiousity. New York: McGraw-Hill.

Bransford, J. D., Goldman, S. R., & Vye, N. J. (1991). Making a difference in people's ability to think: Reflections on a decade of work and some hopes for the future. In R. J. Sternberg & L. Okagaki (Eds.), Influences on children (pp. 147-180). Hillsdale, NJ: Erlbaum.

Britton, B., & Graesser, A.C. (1996)(Eds.) Models of understanding text. Hillsdale, NJ: Erlbaum.

Chi, M. T. H., de Leeuw, N., Chiu, M., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. Cognitive Science, 18, 439-477.

Chinn, C., & Brewer, W. (1993) The role of anomalous data in knowledge acquisition: A theoretical framework and implications for science instruction. Review of Educational Research, 63, 1-49.

Clark (1996). Using language. Cambridge: Cambridge University Press.

Collins, A. (1988). Different goals of inquiry teaching. Questioning Exchange, 2, 39-45.

Costa, P.T., & McCrae, R.R. (1991). NEO: Five Factor Inventory. Odessa, FL: Psychological Assessment Resources.

Coté, N., Goldman, S., & Saul, E.U. (1998). Students making sense of informational text: Relations between procesing and representation. Discourse Processes, 25, 1-53.

Davey, B., & McBride, S. (1986). Effects of question generation on reading comprehension. Journal of Educational Psychology, 78, 256-262.

Department of Defense (1983). Armed Services Vocational Aptitude Battery, Form 12a. Washington, D.C.: Department of Defense.

Dillon, T.J. (1988). Questioning and teaching: A manual of practice. New York: Teachers College Press.

Ericsson, K. A., & Simon, H. A. (1993). Protocol Analysis: Verbal Reports as Data (rev. edn). Cambridge, MA: MIT Press.

Festinger, L. (1957). A theory of cognitive dissonance. Evanston, IL: Row, Peterson.

Fishbein, H.D., Eckart, T., Lauver, E., Van Leeuwen, R., & Langmeyer, D. (1990). Learners' questions and comprehension in a tutoring setting. Journal of Educational Psychology, 82, 163-170.

Flammer, A. (1981). Towards a theory of question asking. Psychological Research, 43, 407-420.

Forbus, K. (1984). Qualitative process theory. Artificial intelligence, 24, 85-168.

Gentner, D., & Stevens, A.L. (1983)(Eds.) Mental models. Hillsdale, NJ: Erlbaum.

Graesser, A.C., Baggett, W., & Williams, K. (1996). Question-driven explanatory reasoning. Applied Cognitive Psychology, 10, S17-S32.

Graesser, A.C., Bertus, E.L. (1998). The construction of causal inferences while reading expository texts on science and technology. Scientific Studies of Reading, 2, 247-269.

Graesser, A.C., Clark, L.F. (1985). Structures and procedures of implicit knowledge. Norwood, N.J.: Ablex.

Graesser, A. C., Gordon, S. E., & Brainerd, L. E. (1992). QUEST: A model of question answering. Computers and Mathematics with Applications, 23, 733-745.

Graesser, A. C. & Hemphill, D. (1991). Question answering in the context of scientific mechanisms. Journal of Memory and Language, 30, 186-209.

Graesser, A. C., Lang, K. L., & Roberts, R. M. (1991). Question answering in the context of stories. Journal of Experimental Psychology: General, 120, 254-277.

Graesser, A.C., & McMahen, C.L. (1993). Anomalous information triggers questions when adults solve quantitative problems and comprehend stories. Journal of Educational Psychology, 85, 136-151.

Graesser, A.C., Millis, K.K., Zwaan, R.A. (1997). Discourse comprehension. Annual Review of Psychology, 48, 163-189.

Graesser, A.C., Olde, B., & Lu, S. (2000, in press). Question-driven explanatory reasoning about devices that malfunction. In T. Filjak (Ed.), Proceedings of the 36th International Applied Military Psychology Symposium.
Graesser, A.C., Olde, B., Pomeroy, V., Whitten, S., Lu, S., & Craig, S. (in press). Inferences and questions in science text comprehension. In book edited by J. Otero and M. Helena (Eds.), Science text comprehension.
Graesser, A. C., & Person, N. K. (1994). Question asking during tutoring. American Educational Research Journal, 31, 104-137.
Graesser, A.C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. Psychological Review, 3, 371-395.
Graesser, A.C., VanLehn, K., Rose, C., Jordan, P., & Harter, D. (in press). Intelligent tutoring systems with conversational dialogue. AI Magazine.
Graesser, A.C., Wiemer-Hastings, P., & Wiemer-Hastings, K. (in press). Constructing inferences and relations during text comprehension. In T. Sanders, J. Schilperoord, & W. Spooren (Eds.), Text representation: Linguistic and psycholinguistic aspects. Benjamins.
Grice, H.P. (1975). Logic and conversation. In P. Cole and J. Morgan (Eds.), Syntax and Semantics 3: Pragmatics. New York: Academic Press.
Hacker, D.J., Dunlosky, J., & Graesser, A.C. (1998)(Eds.). Metacognition in educational theory and practice. Mahwah, NJ: Erlbaum.
Haviland, S.E., Clark, H.H. (1974). What's new? Acquiring new information as a process in comprehension. Journal of Verbal Learning and Verbal Behavior, 13, 512-521.
Hegarty, M., & Just, M.A. (1993). Constructing mental models of machines from text and diagrams. Journal of Memory and Language, 32, 717-742.
Just, M.A., & Captenter, P.A. (1980). A theory of reading: From eye fixations to comprehension. Psychological Review, 87, 329-354.
Just, M.A., & Captenter, P.A. (1992). A capacity theory of comprehension: Individual differences in working memory. Psychological Review, 99, 122-149.
Kass, A. (1992). Question-asking, artificial intelligence, and human creativity. In T. Lauer, E. Peacock, & A.C. Graesser (Eds.), Questions and information systems (pp. 303-360). Hillsdale, NJ: Erlbaum.
Kieras, D.E., & Bovair, S. (1984). The role of a mental model in learning to operate a device. Cognitive Science, 8, 255-274.
King, A. (1989). Effects of self-questioning training on college students' comprehension of lectures. Contemporary Educational Psychology, 14, 366-381.
King, A. (1992). Comparison of self-questioning, summarizing, and notetaking-review as strategies for learning from lectures. American Educational Research Journal, 29, 303-323.
King A. (1994). Guiding knowledge construction in the classroom: Effects of teaching children how to question and how to explain. American Educational Research Journal, 31, 338-368.
Kintsch, W. (1998). Comprehension: A paradigm for cognition. Cambridge: Cambridge University Press.
LaPointe, L.B., & Engle, R.W. (1990). Simple and complex word spans as measures of working memory capacity. Journal of Experimental Psychology: General, 64, 1118-1133.
Lehmann, F. (1992)(Eds.) Semantic networks in artificial intelligence. New York: Pergamon.
Lenat, D.B. (1995). CYC: A large-scale investiment in knowledge infrastructure. Communications of the ACM, 38, 33-38.
Loftus, G.R., & Mackworth, N.H. (1978). Cognitive determinants of fixation location during picture viewing. Journal of Experimental Psychology: Human Perception and Performance, 4, 565-572.
Macaulay, D. (1988). The way things work. Boston: Houghton Mifflin.
Magliano, J.P., Graesser, A.C., Eymard, L.A., Haberlandt, K., & Gholson, B. (1993). The locus of interpretive and inference processes during text comprehension: A comparison of gaze durations and word reading times. Journal of Experimental Psychology: Learning, Memory, and Cognition, 19, 704-709.
Marshall, S. (1999). Cognitive applications of new computational technologies in eye tracking. Invited keynote address at the Artificial Intelligence in Education conference, Le Mans, France.
Mayer, R.E. (1997). Multimedia learning: Are we asking the right questions? Educational Psychologist, 32, 1-19.
Mayer, R.E., & Sims, V.K. (1994). For whom is a picture worth a thousand words? Extensions of a dual-code theory of multimedia learning. Journal of Educational Psychology, 86, 389-401.
Millis, K., & Graesser, A. (1994). The time-course of constructing knowledge-based inferences for scientific texts. Journal of Memory and Language, 33, 583-599.
Miyake, N. & Norman, D.A. (1979). To ask a question one must know enough to know what is not known. Journal of Verbal Learning and Verbal Behavior, 18, 357-364.

O'Brien, E.J., Raney, G.E., Albrecht, J.E., & Rayner, K. (1997). Processes involved in the resolution of anaphors. Discourse Processes, 23, 1-24.
Otero, J., & Campanario, J.M. (1990). Comprehension evaluation and regulation in learning from science texts. Journal of Research in Science Teaching, 27, 447-460.
Otero, J., & Graesser, A.C. (in press). PREG: Elements of a model of question asking. Cognition and Instruction.
Palinscar, A. S., & Brown, A. (1984). Reciprocal teaching of comprehension-fostering and comprehension-monitoring activities. Cognition and Instruction, 1, 117-175.
Papert, S. (1980). Mindstorms: Children, computers, and powerful ideas. New York: Basic Books.
Perfetti, C.A., Britt, M.A., & Georgi (1995). Text-based learning and reasoning: Studies in history. Hillsdale, NJ: Erlbaum.
Piaget, J. (1952). The origins of intelligence. Madison, CT: International Universities Press.
Pressley, M., & Levin, J.R. (1983). Cognitive strategy training: Educational applications. New York: Springer Verlag.
Rayner, K., & Pollatsek, A. (1989). The psychology of reading. Englewood Cliffs, NJ: Prentice-Hall.
Reder, L.M., & Cleeremans, A. (1990). The role of partial matches in comprehension: The Moses illusion revisited. In A.B. Graesser and G.H. Bower (Eds)., The Psychology of Learning and Motivation: Inferences and Text Comprehension (pp. 233-258). San Diego: Academic Press.
Reisbeck, C.K. (1988). Are questions just function calls? Questioning Exchange, 2, 17-24.
Schank, R.C. (1986). Explanation patterns: Understanding mechanically and creatively. Hillsdale, NJ: Erlbaum.
Schank, R.C. (1999). Dynamic memory revisited. Cambridge: Cambridge University Press.
Schank, R.C., Kass, A., & Riesbeck, C.K. (1994). Inside case-based explanation. Hillsdale, NJ: Erlbaum.
Singer, H., & Donlan, D. (1982). Active comprehension: Problem solving schema with question generation for comprehension of complex stories. Reading Research Quarterly, 17, 166-186.
Singer, M. (1994). Discourse inference processes. In M.A. Gernsbacher (Ed.), Handbook of Psycholinguistics (pp. 479-515). San Diego: Academic Press.
Stanovich, K.E., & Cunningham, A.E. (1992). Studying the consequences of literacy within a literate society: The cognitive correlates of print exposure. Memory & Cognition, 20, 51-68.
Trabasso, T. & Van den Broek. P. (1985). Causal thinking and the representation of narrative events. Journal of Memory and Language, 24, 612-630.
Van der Meij, H. (1988). Constraints on question asking in classrooms. Journal of Educational Psychology, 80, 401-405.
VanLehn, K. (1990). Mind bugs: The origins of procedural misconceptions. Cambridge, MA: MIT Press.
VanLehn, K., Jones, R.M., & Chi, M.T. (1992). A model of the self-explanation effect. The Journal of the Learning Sciences, 2, 1-59.
Wiley, J. (1999). Using tasks and browser design to support understanding from web documents. Unpublished manuscript, Washington State University.
Wiley, J. & Voss, J. F. (1999) Constructing arguments from multiple sources: Tasks that promote understanding and not just memory for text. Journal of Educational Psychology, 91, 1-11.
Zimmerman, B.J. (1989). A social cognitive view of self-regulated academic learning. Journal of Educational Psychology, 81, 329-339.

8. List of Publications and Conference Presentations

A. Directly Related to Funded Project

Publications

Craig, S.D., Gholson, B., Ventura, M., Graesser, A.C., & the TRG (in press). Overhearing dialogues and monologues in virtual tutoring sessions: Effects on questioning and vicarious learning. International Journal of Artificial Intelligence in Education.

Graesser, A.C., Olde, B., & Lu, S. (2000, in press). Question-driven explanatory reasoning about devices that malfunction. In T. Filjak (Ed.), Proceedings of the 36th International Applied Military Psychology Symposium.

Graesser, A.C., Olde, B., Pomeroy, V., Whitten, S., Lu, S., & Craig, S. (in press). Inferences and questions in science text comprehension. In book edited by J. Otero and M. Helena (Eds.), Science text comprehension.

Otero, J., & Graesser, A.C. (in press). PREG: Elements of a model of question asking. Cognition and Instruction.

Conference Presentations (chronological order)

Graesser, A.C., Pomeroy, V., & Craig, S. (1999, January). Think aloud protocols in the context of illustrated texts and devices that malfunction. Paper presented at the Tenth Annual Meeting of the Winter Conference on Discourse, Text & Cognition, Jackson Hole, Wyoming.

Graesser, A.C. (1999, February). Question-driven explanatory reasoning about devices that malfunction. Progress report to the Navy on Office of Naval Research grant, Oxford, Mississippi.

Graesser, A.C. (1999, February). Deep comprehension of illustrated texts about everyday devices. Colloquium at the University of Coimbra, Coimbra, Portugal.

Graesser, A.C. (1999, February). Question-driven explanatory reasoning about devices that malfunction. Progress report to the Navy on Office of Naval Research grant, Millington, Tennessee.

Graesser, A.C., Craig, S., Pomeroy, V., & Olde, B. (1999, April). Deep comprehension of illustrated texts in the context of a breakdown scenario. Invited symposium on discourse comprehension at the meetings of the American Educational Research Association, Montreal, Canada.

Graesser, A.C., Craig, S., Pomeroy, V. & Olde, B. (1999, July). Comprehension of illustrated texts about everyday devices. Paper presented at the meetings of the Society for Applied Research in Memory and Cognition, Boulder, Colorado.

Graesser, A.C., Whitten, S.N., & Olde, B. (June, 2000). Do males and females differ on mechanical comprehension? Poster presented at the meetings of the American Psychological Society, Miami, FL.

Graesser, A.C., Lu, S., Whitten, S., Olde, B., Pomeroy, V., & Craig, S. (July, 2000). Deep and shallow comprehension of illustrated texts on everyday devices. Paper presented at the meetings of the Society for the Scientific Studies of Reading, Stockholm, Sweden.

Graesser, A.C., Lu, S., & Whitten, S. (July, 2000). Inferences about causal mechanisms that are depicted in illustrated texts on everyday devices. Paper presented at the International Congress of Psychology, Stockholm, Sweden.

Whitten, S.N., Lu, S., & Graesser, A.C. (July, 2000). What determines deep comprehension for illustrated texts. Poster presented at the meetings of the Society for Text and Discourse, Lyon, France.

Graesser, A.C. (September, 2000). Question-driven explanatory reasoning about devices that malfunction. 36th International Applied Military Psychology Symposium, Split, Croatia.

B. Publications Indirectly Related to Funded Project

Corbett, A., Anderson, J., Graesser, A., Koedinger, K., & van Lehn, K. (1999). Third generation computer tutors: Learn from or ignore human tutors? Proceedings of the 1999 Conference of Computer-Human Interaction (pp. 85-86). New York: ACM Press.

Du Boulay, B., Greer, J., Lepper, M., Graesser, A., Van Lehn, K., & Moore, J. (1999). Panel on issues involving human and computer tutoring. In S.P. Lajoie and M. Vivet, Artificial Intelligence in Education (pp. 780-781). Amsterdam: IOS Press.

DiPaolo, R.E., Graesser, A.C., Hacker, D.J., White, H.A., & TRG (in press). Hints in human and computer tutoring. In M. Rabinowitz (Ed.), The impact of media on technology of instruction. Mahwah, NJ: Erlbaum.

Graesser, A.C., & Bertus, E.L. (1998). The construction of causal inferences while reading expository texts on science and technology. Scientific Studies of Reading, 2, 247-269.

Graesser, A.C., Franklin, S., & Wiemer-Hastings, P. and the Tutoring Research Group (1998). Simulating smooth tutorial dialogue with pedagogical value. Proceedings of the American Association for Artificial Intelligence (pp. 163-167). Menlo Park, CA: AAAI Press.

Graesser, A.C., Karnavat, A., Pomeroy, V., Wiemer-Hastings, & TRG (2000). Latent semantic analysis captures vestiges of causal, goal-oriented, and taxonomic structures. Proceedings of the Cognitive Science Society.

Graesser, A.C., Kassler, M.A., Kreuz, R.J., & McLain-Allen, B. (1998). Verification of statements about story worlds that deviate from normal conceptions of time: What is true about *Einstein's Dreams*? Cognitive Psychology, 35, 246-301.

Graesser, A.C., Kennedy, T., Wiemer-Hastings, P., & Ottati, V. (1999). The use of computational cognitive models to improve questions on surveys and questionnaires. In M.G. Sirken, D.J. Hermann, S. Schechter, N. Schwarz, J.M. Tanur, & R. Tourangeau (Eds.), Cognition and survey methods research (pp 69-86). New York: Wiley.

Graesser, A.C., Person, N., Harter, D., & TRG (2000). Teaching tactics in AutoTutor. Proceedings of the workshop on modeling human teaching tactics and strategies at the Intelligent Tutoring Systems 2000 conference. University of Quebec at Montreal, 49-57.

Graesser, A.C., VanLehn, K., Rose, C., Jordan, P., & Harter, D. (in press). Intelligent tutoring systems with conversational dialogue. AI Magazine.

Graesser, A.C., Wiemer-Hastings, P., & Wiemer-Hastings, K. (in press). Constructing inferences and relations during text comprehension. In T.Sanders, J. Schilperoord, & W. Spooren (Eds.), Text representation: Linguistic and psycholinguistic aspects. Amsterdam: Benjamins.

Graesser, A.C., Wiemer-Hastings, K., Wiemer-Hastings, P., Kreuz, R., & TRG (1999). AutoTutor: A simulation of a human tutor. Journal of Cognitive Systems Research, 1, 35-51.

Graesser, A.C., Wiemer-Hastings, K., Kreuz, R., & Wiemer-Hastings, P. (2000). QUAID: A questionnaire evaluation aid for survey methodologists. Behavior Research Methods, Instruments, and Computers, 32, 254-262.

Graesser, A.C., Wiemer-Hastings, K., Wiemer-Hastings, P., & Kreuz, R. (2000). The gold standard of question quality on surveys: Experts, computer tools, versus statistical indices. Proceedings of the American Statistical Association.

Graesser, A.C., Wiemer-Hastings, P., Wiemer-Hastings, K., Harter, D., Person, N., and the TRG (2000). Using latent semantic analysis to evaluate the contributions of students in AutoTutor. Interactive Learning Environments, 8, 74-103.

Hacker, D.J., Dunlosky, J., & Graesser, A.C. (1998)(Eds.). Metacognition in educational theory and practice. Mahwah, NJ: Erlbaum.

Hu, X., Graesser, A.C., and the Tutoring Research Group (1998). Using WordNet and latent semantic analysis to evaluate the conversational contributions of learners in tutorial dialog. Proceedings of the International Conference on Computers in Education, Vol. 2 (pp. 337-341). Beijing, China: Springer.

Magliano, J., Trabasso, T., & Graesser, A.C. (in press). Strategic processing during comprehension. Journal of Educational Psychology.

McCauley, L., Gholson, B., Hu, X., Graesser, A.C., and the Tutoring Research Group (1998). Delivering smooth tutorial dialogue using a talking head. Proceedings of the Workshop on Embodied Conversation Characters (pp. 31-38). Tahoe City, CA: AAAI and ACM.

Person, N.H., Bautista, L., Kreuz, R.J., Graesser, A.C., & TRG (June, 2000). The Dialog Advancer Network: A Conversation Manager for AutoTutor. Proceedings of the workshop on modeling human teaching tactics and strategies at the Intelligent Tutoring Systems 2000 conference. University of Quebec at Montreal, 86-92.

Person, N.K., Craig, S., Price, P., Hu, X., Gholson, B., Graesser, A., & TRG (2000). Incorporating human-like conversational behaviors in AutoTutor. Proceedings of the Workshop on Conversational Agents at the Fourth International Conference on Autonomous Agents 2000 (pp. 85-92). Barcelona, Spain: ACM Press.

Person, N.K, & Graesser, A.C. (1999). Evolution of discourse in cross-age tutoring. In A.M. O'Donnell and A. King (Eds.), Cognitive perspectives on peer learning (pp.69-86). Mahwah, NJ: Erlbaum.

Person, N.K., Graesser, A.C., & TRG (2000). AutoTutor's conversational behaviors. Proceedings of the Third Workshop on Human-Computer Conversation.

Person, N.K., Graesser, A.C., and the Tutoring Research Group (2000). Designing AutoTutor to be an effective conversational partner. Proceedings of the Fourth International Conference of the Learning Sciences.

Person, N.K., Graesser, A.C., Harter, D., Mathews, E., & TRG (2000). Dialog move generation and conversation management in AutoTutor. Proceedings of the AAAI Fall Symposium 2000 on Building Dialogue Systems for Tutorial Applications (pp 45-51). Menlo Park, CA: AAAI Press.

Person, N.K., Graesser, A.C., Kreuz, R.J., Pomeroy, V., & TRG (in press). Simulating human tutor dialog moves in AutoTutor. International Journal of Artificial Intelligence in Education.

Wiemer-Hastings, P., Graesser, A.C., Harter, D., and the Tutoring Research Group (1998). The foundations and architecture of AutoTutor. Proceedings of the 4th International Conference on Intelligent Tutoring Systems (pp. 334-343). Berlin, Germany: Springer-Verlag.

Wiemer-Hastings, P., Wiemer-Hastings, K., Graesser, A., and the Tutoring Research Group (1999). Approximate natural language understanding for an intelligent tutor. Proceedings of the American Association for Artificial Intelligence (pp. Xx-xx). Menlo Park, CA: AAAI Press.

Wiemer-Hastings, P., Wiemer-Hastings, K., and Graesser, A. (1999). Improving an intelligent tutor's comprehension of students with Latent Semantic Analysis. In S.P. Lajoie and M. Vivet, Artificial Intelligence in Education (pp. 535-542). Amsterdam: IOS Press.

Wiemer-Hastings, K., Wiemer-Hastings, P., Rajan, S., Graesser, A.C., Kreuz, R.J., & Karnavat, A. (2000). DP—A detector for presuppositions in survey questions. Proceedings of the Joint Language Technology Conference (pp. 90-96). ACL Press.

Williams, K.E., Hultman, E., & Graesser, A.C. (1998). CAT: A tool for eliciting knowledge on how to perform procedures. Behavior Research Methods, Instruments, & Computers, 30,565-572.

9. Students Working on this Funded Research

Doctoral Students

Scotty Craig
Shulan Lu
Brent Olde
Shannon Whitten

Masters Students

Victoria Pomeroy

Undergraduate Students

Elisa Cooper
Frances Daniel